QROWD - Because Big Data Integration is Humanly Possible

Innovation Action

Grant agreement no.: 732194

D10.1 – Data Management Plan (Update)

Due Date	30 Nov 2019
Actual Delivery Date	30 Nov 2019
Document Author/s	Luis-Daniel Ibáñez (University of Southampton)
Version	2.0
Dissemination level	PU
Status	Final / Update Provided after the end of the project

Project co-funded under Horizon 2020 Research and Innovation EU programme

TABLE OF CONTENT

LIST OF ABBREVIATIONS

(if there is need for such)

FAIR	Findable, Accessible, Interoperable and Re-usable
DMP	Data Management Plan
OASC	Open Access Smart Cities Initiative

EXECUTIVE SUMMARY

The Data Management Plan (DMP) describes QROWD's data management life cycle for the data to be collected, processed and/or generated, as part of making its research data findable, accessible, interoperable and re-usable (FAIR), following the guidelines for FAIR Data Management[1]. FAIR data management guarantees that the advancements and results developed on top of these data can be replicated and exploited by future EU-funded initiatives and the community in general. This document provides an update on our processes after the end of the project.

The key target readers of the DMP are the warrants of the Open Research Data Pilot program that will check the compliance of the project with H2020 guidelines, and members of the research community interested in replicating and/or reproducing the results of QROWD's research and development. Technical partners of the consortium, that will use it as a guide for publishing and maintaining data associated to bespoke research.

In this deliverable we describe the processes, policies and tools the QROWD consortium used to ensure FAIR data management, that can be summarized as follows:

- Findability: Datasets produced by research in the project will be assigned a DOI and deposited in Zenodo. Datasets meant to be part of the open source
- Accessibility: Research datasets will be published in Zenodo and linked to OpenAire
- Interoperability: In addition to the descriptions used for accessibility, we will select the most appropriate standards to format the data. We will pay particular attention to those issued from the OASC initiative.
- Re-usability: All research data outputs will be published with an open license with the exception of those corresponding to the TomTom use case.

We also describe the measures we took regarding data protection and privacy with personal data.

The deliverable is complemented by the data catalog (D4.1), that was also updated to reflect changes in the second half of the project.

[1] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

SUMMARY OF UPDATES

- On Data Summary was updated to reflect updates on the Data Catalog (D4.1), that now features the list of datasets that were used during the whole project (It was previously up to M12)
- On Data Summary, update on the non-publication of datasets including GPS coordinates of citizens, as deemed personal data with publication outside of the scope of consent given
- On FAIR, we removed mentions to DCAT-AP extension for scientific datasets to describe our datasets, as it stayed only as draft. We added mentions to our usage of ML-Schema, and crowd-voc (described in D3.4)
- On FAIR, we now state our choice of Fiware Data Models to increase interoperability with cities that have adopted OASC standards.
- Throughout, we added Zenodo as the main archive for data and software outputs.
- Added a section on Data Protection, describing the assessment of the risks and measures we took for processing the personal data required for the modal split use case.
- Add mention to data we did not publish due to Data Protection considerations: GPS stream and trip confirmation data.

1 DATA SUMMARY

Data in QROWD is divided in three groups, according to its use and its relation with other WPs:

1. Data collected and generated for testing and improving analytics and hybrid discovery capabilities of the QROWD's platform. This comprises the input data in RDF format on which the analysis/discovery processes will be run, the "ground truth" data used to assess the effectiveness of the newly developed approaches, and the crowdsourced data used either as part of the "ground truth", or to improve the analysis/discovery processes. All these data will be made openly accessible.
2. Data collected and generated for the Trento municipality use case (cf. D2.2). This comprises data transformed from municipality's data sources (static datasets and sensor data) and data collected from citizens as a complement to sensor data. Transformations from Open Data will remain open. In the initial version of this document, we expected that an appropriately curated and anonymized subset of the crowdsourced data that allows the reproduction of experiments conducted in connection with this use case will be published, provided that data subjects have agreed with the release of their particular GPS traces. However, GPS traces were considered sensitive data by the Data Protection Officer, and consent was only granted for a limited time (up to the end of the project). Therefore, we decided to not make GPS traces openly available. Only
3. Data collected and generated for the TomTom use case (cf D1.1). This use case re-uses some data from the Trento municipality, but also includes data proprietary to TomTom that cannot be released due to being core to TomTom's business model.

During the project, Data will be under the custody of the research/industrial partner leading the work package, in the case of WP2, UniTN and MT will be joint controllers. In the case of WP1, the custody is shared between TomTom and InfAI. During this phase, access to the data will be restricted to consortium partners.

The full list of data used and collected in the project is in the Data Catalog deliverable (c.f. D4.1)). Most of it is data already considered as open and without any GDPR implications, except for a combination of identifying fields, and sensor streams collected through the i-Log application, namely

- Name, email, gender, age range, number of vehicles available, number of members of the household, preferred vehicle
- GPS stream
- Accelerometer stream
- Gyroscope stream
- User feedback from inferred trips

- User input for missing trips

Section 5 describes the data protection measures we took to guarantee a seamless and GDPR compliant workflow.

2 FAIR DATA

2.1 FINDABILITY

Each dataset that we consider important for research purposes developed during the project will be assigned a Digital Object Identifier through the Zenodo service. In the initial version of this document we considered the use of the DCAT-AP extension for scientific datasets recently proposed by the European Commission Joint Research Centre[2]. However, as the DCAT-AP extension for scientific datasets did not proceed further than unofficial draft, we ultimately decided to only use Zenodo.

Datasets acquired and transformed through the data acquisition framework (D4.2) will be findable through the internal CKAN infrastructure of the framework for use by the project consortium.

Internal versioning will follow the incremental 0.x format until publication of the linked scientific contribution, point from which the numeration will follow the 1.x format for minor corrections or improvements.

2.2 ACCESSIBILITY

Data corresponding to research conducted for the Trento municipality use case considered as anonymous and or securely pseudonymised will be released (c.f. section 5). Data corresponding to the TomTom use case will not be publicly available due to the reliance of the business model of TomTom on it. The possibility of releasing a subset of the data is currently being discussed internally. This plan will be updated according to the final decision made.

Code corresponding to research conducted for WP5 and WP6 will be released to the community with an open license.

While research is being undertaken, accessibility to datasets will be limited to the members of the consortium. During this stage, partners leading the specific research will be in charge of the storage and accessibility for the rest of the

[2] https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_27

partners that require it.

Following the directives of OpenAire, all research outputs that could be released with open licenses are deposited in Zenodo[3]. This includes releases from QROWD's Github repository corresponding to the latest versions of code used in the project.

Concerning the tools required to access QROWD's research data. At the time of the first version of this plan, we initially expected that all datasets will be available in RDF format. Any RDF Graph-Store and SPARQL engine can be used to load them and query them. As a result of the re-use of other established data models and formats (e.g. those used by OASC), only a fraction of the produced datasets are in RDF.

2.3 INTEROPERABILITY

To ensure inter-operability, we aligned our datasets to the FiWare data models for transportation and parking[4], as it is mostly used by the OASC organisation. We also re-used the ML-Schema vocabulary to describe datasets resulting from We also developed the crowd-voc[5] vocabulary for describing datasets produced with Crowdsourcing.

2.4 RE-USABILITY

Datasets connected to the Trento Municipality use-case (WP2) will be re-usable according to one of the following schemes

1. Data transformed from existing data sources (curated or not) will have the same license as the original source.
2. Data collected and generated from the project that is connected to a scientific publication will be made available with a CreativeCommons license. In the case of data coming from crowdsourcing, appropriate anonymisation processes will be apply before release (cf. D11.1).

Datasets connected to the TomTom use-case (WP1) will not be re-usable.

3 RESOURCE ALLOCATION

The archiving infrastructure and resources will be provided by Southampton and UniTN. Successive updates of the DMP will be led by Soton, as coordinating partner.

[3] https://zenodo.org/search?page=1&size=20&q=qrowd
[4] https://www.fiware.org/developers/data-models/
[5] https://doi.org/10.5281/zenodo.3373397

4 DATA SECURITY

Southampton, has a secure enterprise scale coherent storage solution for active research data. The data stored within this facility is regularly backed up and a copy of the back-up, regularly off-sited to a secure location for disaster recovery purposes. The research data storage platform is solely for the storage of research data. Final versions of datasets will be deposited in Zenodo.
UniTN's infrastructure abides to the European Commission Recommendation on access to and preservation of scientific information (July 17th 2012) the H2020/ERC Model Grant Agreement, that has all the security measures to avoid data loss and intrusion.

Data for the TomTom business case will be hold by TomTom, following their industry-grade security measures. Concerning location-identifying data, we cite the following excerpt from TomTom's policy:

"*Within 24 hours of you shutting down your device or app, TomTom automatically and irreversibly destroys the data that would allow you or your device to be identified from the location data we received.*

For Traffic, SpeedCameras, Danger Zones and Weather we delete the information within 20 minutes after you have stopped using the service by shutting down your device or app. We do not know where you have been and cannot tell anyone else, even if we somehow were forced to.

This, now anonymous, information is used to improve TomTom's products and services, such as TomTom maps, Traffic, products based on traffic patterns and average speeds driven, and for search queries to inform businesses how well-received their information is. These products and services are also used by government agencies and businesses."

TomTom data used in the project is always aggregated or stripped of its identifiers.

5 DATA PROTECTION

QROWD processes personal data of citizens as part of the modal split estimation of WP2. In this section, we detail the measures we took for ensuring GDPR compliance, on the light of the need for several partners of the consortium to process data.

Under the advice and assistant of the Data Protection Officer of the University of Trento, we carried out a checklist to assess data protection. The original document, in Italian, is annexed to this document. We summarize the key details as follows.

Roles were established as follows:

- Joint Data Controllers: Municipality of Trento and University of Trento
- Data Processors: University of Southampton, InfAI

The following table summarizes the data collected and who processed it.

Data field **(P = Personal)**	**Purpose**	**Processed by**
Name (P)	To address citizen	Controllers
Email (P)	Contact citizen	Controllers
Gender (P)	Enable aggregations by gender	Controllers
Age range	Enable aggregations by age range	Controllers
Vehicles available and preferred	Enable aggregations	Controllers
GPS Trace (P)	Automatic detection of trips and changes of transport mode.	Controllers and processors
Accelerometer	Automatic detection of trips and changes of transport mode.	Controllers and processors
Gyroscope	Automatic detection of trips and changes of transport mode.	Controllers and processors
Manual confirmation and input of trips (P)	Validate automated detection of trips and changes of transport mode. Collect trip data	Controllers and processors
Preferred time to receive questions	Sets time to send questions to users.	Controllers and processors

The evaluation revealed that none of the conditions of article 35.3 apply to our collection and processing, namely

- No decisions with legal effect will be taken based on the collected data
- No special categories of personal data involved
- No large scale monitoring of public areas

Inline with the principle of minimisation. a pseudonymous was generated by controllers to identify traces belonging to the same user, so the identifying fields name and email would not need to be accessible to data processors. We also designed the API calls available to data processors to avoid any data leakage with respect to demographic information, returning only aggregated information.

All identified risks on confidentiality, integrity, and availability, were evaluated as "Low" or "Very Low". The table below shows the risk log.

Risk #	Description	Probability	Impact	Mitigation
1	Lost of device containing personal data	Low	High	Minimize the number of devices where personal data is stored
2	Personal data wrongly sent to unauthorized party	Low	Medium	Check for record integrity before sending personal data back to participants
3	Web server/service misconfiguration leaks personal data	Very low	High	Audit server configurations before experiments
4	Information (e.g. Google accounts) of a participant needed for providing a service is modified	Low	Low	Instruct participants to keep records stable during the experiments
5	A linking record is lost.	Low	Low	Correct backup management
6	Lost of data of a participant	Low	Medium	Correct backup management
7	A critical service is down	Very low	Medium	Infrastructure test.